(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
18 October 2001 (18.10.2001)

PCT

(10) International Publication Number
WO 01/77381 A2

(51) International Patent Classification[7]: C12Q 1/68

(21) International Application Number: PCT/GB01/01633

(22) International Filing Date: 11 April 2001 (11.04.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
0008899.7 11 April 2000 (11.04.2000) GB

(71) Applicant *(for all designated States except US)*: **ISIS INNOVATION LIMITED** [GB/GB]; Ewert House, Ewert Place, Summertown, Oxford OX2 7SG (GB).

(72) Inventor; and

(75) Inventor/Applicant *(for US only)*: **SYKES, Bryan, Clifford** [GB/GB]; Wolfson College, Oxford, Oxfordshire OX2 6UD (GB).

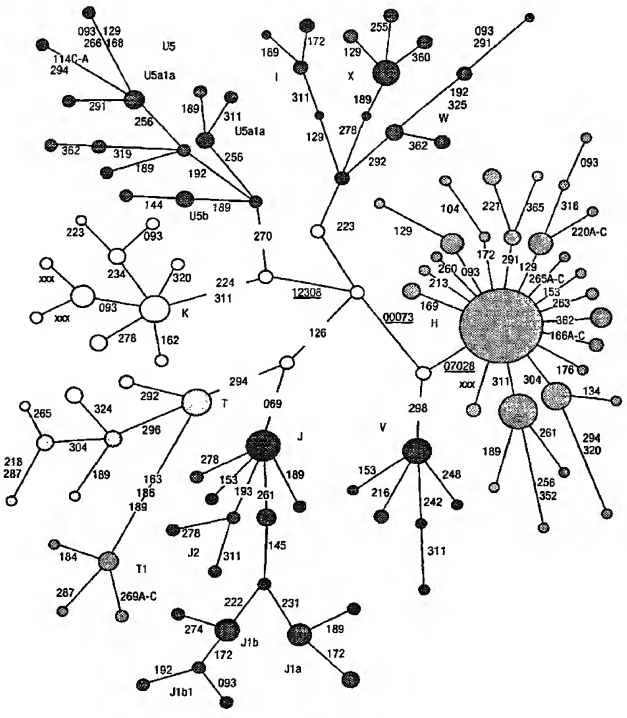
(74) Agents: **WRIGHT, Simon, Mark** et al.; J.A. Kemp & Co, 14 South Square, Gray's Inn, London WC1R 5LX (GB).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

[Continued on next page]

(54) Title: DNA ANALYSIS



mitochondrial DNA sequence

ATTCTAATTT AAACTATTCT CTGTTCTTTC ATGGGGAAGC AGATTTGGGT ACCACCCAAG TATTGACTCA CCCATCAACA ACCGCTATGT ATTTCGTACA
TTACTGCCAG CCACCATGAA TATTGTACGG TACCATAAAT ACTTGACCAC CTGTAGTACA TAAAAACCCA ATCCACATCA AAACCCCCTC CCCATGCTTA
CAAGCAAGTA CAGCAATCAA CCCCCAACTA TCACACATCA ACTGCAACTC CAAAGCCACC CCTCACCCAC TAGGATACCA ACAAACCTAC CCACCCTTAA
CAGTACATAG TACATAAAGC CATTTACCGT ACATAGCACA TTACAGTCAA ATCCCTTCTC GTCCCCATGG ATGACCCCCC TCAGATAGGG GTCCCTTGAC

(57) Abstract: A method of determining probable ancestry of a human individual to a dead human female ancestor is disclosed. A sample is taken from a human whose ancestry is to be determined, mitochondial DNA (mtDNA) extracted therefrom and sequenced. A portion of this mtDNA sequence is then compared with other "comparison" mtDNA sequences in the same region to determine the differences (or mutations) between these two sequences. In particular, the sequence from a human is compared with a reference sequence and the positions of the mutations determined. Sequences having the same mutations are likely to be related, and thus an indication of probable ancestryy can be determined. A database of sequences can be produced, where the sequences are related to each other by means of their mutations, and this can be used to generate a gene tree. An individuals' position in that gene tree, determined by the nature of the mutation, can then indicate ancestry, in particular to a female ancestor who lived many thousands of years ago. In particular, it has been found that virtually all human mtDNA sequences can be traced back to only a few (6-8) female ancestors.

WO 01/77381 A2

Published:
— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

DNA ANALYSIS

The present application concerns a method of determining probable ancestry, in particular relation to a human female ancestor, by use of sequence information from mitochondrial DNA sequences.

DNA analysis and comparison of DNA sequences have been used for many years to provide information about those sequences. For example, if one discovers a new sequence, then information concerning that sequence can be obtained by comparing the new sequence with known sequences, and from the function and/or purpose of those known sequences one may be able to glean some information about the corresponding function and purpose of the new sequence. However, when it comes to determination of ancestry, or genealogy, there is at present no method for determining potential ancestors by comparing DNA sequences. Although some DNA comparisons have been conducted in order to indicate race or geographical origin, at present there has been no use of DNA techniques to indicate potential relation with an ancestor. The present invention sets out to advance genealogy techniques by using mitochondrial DNA sequence analysis.

It has been found that most human mitochondrial sequences fall into one of a number (possibly seven) clades. This allows the clade of an individual to be determined using reference sequences from known clades.

The present invention therefore, at its broadest, relates to a method of determining a probable ancestor of a human, wherein mitochondrial DNA (mtDNA) from the human is compared with mitochondrial DNA sequences, for example in a database, those sequences being correlated to one or more female ancestor(s).

Accordingly the first aspect of the present invention relates to a method of determining probable ancestry, the method comprising:

(a) providing a mitochondrial DNA sequence from a human (the "sample sequence");

(b) comparing the sample sequence with a multiplicity of mitochondrial DNA sequences ("comparison sequences") each from a human different from the human having the sample sequence; and

(c) providing, on the basis of the comparison, an indication that the human having the sample sequence is a probable ancestor of a human female having a related comparison sequence.

The invention in a second aspect relates to a method of determining the clade of an individual, the method comprising:

(a) providing a mitochondrial DNA sequence from a human (the "sample sequence");

(b) comparing the sample sequence with a multiplicity of mtDNA sequences from different clades ("comparison sequences") and determining the closest comparison sequence to the sample sequence; and

(c) determining the clade of the individual, this being the same clade as the closest sequence.

A third aspect of the present invention relates to the method of determining probable ancestry, or obtaining ancestry information, or determining a clade of an individual (human), the method comprising:

(a) providing an mtDNA sequence from a (human) individual (the "sample sequence");

(b) comparing the sample sequence with a reference mtDNA sequence from a human, and determining the differences between the sample and reference sequence (e.g. the mutations);

(c) using the nature of the mutations (e.g. the position (number) and/or the nature of the base substituted for the original) to correlate the sample sequence with a comparison sequence (from a human other than the one with the sample sequence), wherein the sample and comparison sequences both have at least one mutation in common. This can indicate that the human having the sample sequence is related to, is ancestor of, or belongs to the same clade as, a human female having that comparison sequence; and

(d) optionally, providing on the basis of the correlation between the sample sequence and the comparison sequence, a link between a characteristic of the human female ancestor having the comparison sequence to the human having the sample sequence or to determine information about the human female ancestor,

determination of the clade of the individual or a probable ancestor.

The human possessing the sample sequence is preferably alive. They may be male or female. The comparison sequence(s) may belong to dead or alive male or female humans. Preferably, the comparison comprises identification of the closest sequence to the sample sequence. The closest sequence may be from a human who is dead. Preferably the comparison will allow a link with or determination of a single female ancestor. The ancestor may thus have the closest sequence or a (closely) related sequence and may be dead or alive.

Mitochondrial DNA is maternally inherited, fast evolving and non-recombining. This means that by comparing mtDNA sequences one can establish female lineage, or one or more female ancestors, since mitochondrial DNA is passed on only through the egg, and is therefore always from the mother.

The comparison preferably comprises comparing the sample sequence with a comparison sequence that is a reference sequence. This is to determine the difference(s) between the sample and the reference sequence (mutations). The same process may be conducted to determine any common mutations shared by both the sample sequence and a comparison sequence (for example to determine the closest sequence). One may then use the mutations to indicate a probable (female) ancestor, the ancestor having one or more of the same mutations. Thus if there are mutations that are common in the sample and reference sequences this can indicate ancestry. This may involve determining the location and nature of the mutations.

There are two characteristics of any mutation. These are its position, and the mutation itself (there are four bases in DNA, and therefore in theory a mutation one of those four bases can be replaced by one of the other three bases). All the mutations referred to herein are transversions (change of purine base for another purine e.g. C→T, T→C, A→G or G→A) unless otherwise specified. Usually it is the position of the mutation that is determined. The method of the invention may comprise, preferably before (a), taking a tissue sample from a human, extracting mtDNA from that sample, and determining sequence of at least part of that DNA.

The method may also comprise correlating a number of the comparison sequences to each other by comparing them to a reference sequence. One then may determine the differences between the comparison and reference sequences (mutations, and

in particular the position of those mutations). A mutation may then be used to correlate sequences that possess the same mutation. If a sample sequence possesses one or more of the same mutations as a comparison sequence, then the sequences may be of the same clade, or this can indicate ancestry.

Preferably all the comparison sequences are sequences from a individual, rather than consensus sequences.

This correlation can then be used to generate relationships between comparison sequences on the basis of the (same) mutations, for example to generate a "gene tree". One may thus be able to relate the sample sequence to one or more comparison sequences that possess one or more of the mutations. The sample sequence may therefore be correlated to the comparison or reference sequences according to the nature of the mutations (or the difference(s) between the sample and reference sequences). This may allow the sample sequence to be placed on the gene tree and so give an indication of ancestry.

In the prior art, the bottom or root of the gene tree (or diagram) is usually one individual. However, in the present invention, the root of the tree can result in a multiplicity of sequences from different individuals. This root of the tree can be thought of as the original ancestor. It may be possible to relate any sample sequence with one of these original ancestors. The number of ancestors can vary depending on the origin of the comparison sequences. For example, there may be from 6 to 8, 5 to 9, or 3 to 10 roots of the gene tree or original ancestors. This may be the case if the comparison sequences are taken from individuals in Europe. However, the number of roots of the tree or original ancestors can increase, if for example the comparison sequences include the sequences from the entire world, in other words if one was comparing the sample sequence with worldwide sequences. In this case, the number of original ancestors may be from 30 to 35, such as from 25 to 40, optimally from 20 to 50.

The method of the invention may further comprise determining only one single human female ancestor to the human possessing the sample sequence. This ancestor may be the root of the tree, or may be an original ancestor as previously discussed. Although there may be a multiplicity of ancestors, in this step the human having the sample sequence is correlated to only one of these original ancestors, but not to any of the others. The or each of these ancestors are preferably female. They may have died at least 3,000, 5,000, 7,000 or 10,000 years ago. The or each ancestor may have approximately 8 mutations

from the sample sequence, although the number of mutations can vary from 7 to 9, such as from 6 to 10, optimally from 3 to 12 mutations. As the mutation rate can be, on average, one every 20,000 years, it will be seen that these figures provide the possible minimum ages of the ancestors. Thus, the original ancestors may therefore be at least 100,000, 150,000 or at least 200,000 years old.

The method of the invention may further comprise the step of providing a physical output. This may be achieved using a computer system as discussed later. The output may be a printed or visual indication as discussed in the various aspects of the invention. The indication may be a printout, text, gene tree or diagram, or a display, for example on a computer. The information displayed may be of the form corresponding to Figures 1 or 2 of this specification.

The invention in a fourth aspect relates to a method of determining the amount or period of time that a sample sequence has been present in a population in Europe, or the age of a European ancestor, the method comprising:

(a) comparing the sample sequence to a multiplicity of mtDNA sequences of a Near East population to determine the closest Near East sequence; and

(b) determining the age of the sample sequence or ancestor in Europe based on the number of mutations between the candidate sequence and the closest Near East sequence.

Once the number of mutations between the two (sample and closest) sequences have been determined, this number can then be multiplied by the determined mutation rate (for example 20,180 years) to determine the time or age as above.

The method of the invention, in particular the comparison, may comprise determining the differences between the sample sequence and a reference sequence, in other words determining the mutations. This may involve determining the position of the mutations and/or the nature of the mutation (that is to say, the base which has been substituted for original base). The position of the mutation is usually more important, and therefore the comparison preferably comprises determining the position(s) of the mutations that the sample sequence has relative to the reference sequence.

It has been found that mutations in certain positions are more frequent that mutations in other positions. Furthermore, the position of the mutation may provide an indication of ancestry (for example clade, group or clan). The invention may thus

comprise determining whether the sample sequence possesses a mutation at one of more of these positions. The positions contemplated include 298, 126 (preferably with 294 and/or 069), 224 and/or 311, 270 and/or 223. If a sample sequence has a mutation at one or more of these positions then this may indicate that the individual having the sample sequence is related to one or more clades or clans. These positions correspond to the section of mtDNA known as HVS I, minus 16,000. In other words, position 069 corresponds to the base at position 16,069 in the region HVS I.

The invention may thus involve determining whether the sample sequence has a mutation at any of these positions, and this may be achieved by using one or more primers or probes that are specific for that mutation, or for a mutation at that position. In particular, the primer or probe should be able to detect a mutation at any particular desired position, for example the position of one of the mutations mentioned above.

A probe may be designed so that it will hybridise, under chosen conditions, to a portion of the sample mtDNA sequence that includes the position at which a mutation is desired to be detected. Several probes may be provided if it is decided to detect mutations at one or more positions.

Primers may be employed to amplify a particular region of mtDNA, for example using PCR. Like the probes, the primers may be, for example, from 7 to 20 bases in length, such as from 10 to 15 bases. The primers may be used to amplify the entire region of mtDNA of interest (such as HVS I) or a part within this region. For example, it may be desired to amplify a portion of the sequence that includes one of the positions mentioned above that is suspected of having a mutation. In other words, one can amplify a portion of mtDNA either side one of the positions where it is desired to detect whether or not there is a mutation. This portion may be up to 5, for example up to 10, bases either side of the position. The present invention therefore includes one or more primers and/or probes that are specific for detecting a mutation at one or more of the positions mentioned above. These primers and/or probes may be present in a kit, and so a further aspect of the present invention relates to a kit comprising one or more probes and/or primers of the invention.

In this specification the genealogy or ancestry includes pedigree, family tree, line, lineage, descent, parentage, family, dynasty, house, bloodline, heritage or history.

The method of the invention, in particular the comparison, may comprise determining the differences between the sample sequence and a reference sequence, in

other words determining the mutations. This may involve determining the position of the mutations and/or the nature of the mutation (that is to say, the base which has been substituted for original base). The position of the mutation is usually more important, and therefore the comparison preferably comprises determining the position(s) of the mutations that the sample sequence has relative to the reference sequence.

A fifth aspect of the present invention relates to a database comprising at least 10 mtDNA sequences, the database being structured so that each sequence is correlated with one or more pieces of information concerning a human female ancestor having that DNA sequence. The information may concern genealogy, for example the age of the ancestor, the geographical origin or location or region of the ancestor, the ancestor's race, clan, clade or grouping, or lifestyle information of the ancestor. This database may be present on a data carrier, and so this aspect additionally includes a data carrier possessing the database.

A sixth aspect of the invention relate to a use of the database of the third aspect in the determination of a probable human female ancestor or clade from a mitochondrial DNA sequence.

The database may have a number of "daughter" sequences (for example 3 to 10, such as 5 to 9, preferably from 6 to 8, and ultimately 7). These daughter sequences may have from 1 to 4 (or 3) differences (or mutations) from a reference sequence. Most of the comparison sequences (at least 50, 60, 70, 80 or 90 or even at least 95 or at least 99% of the comparison sequences in the database) possess one or more of the same mutations (with respect to the reference sequence). This allows sequences to be structured on the database according to the nature of the mutations. Sequences having one or more of the same mutations may be so indicated or correlated, as they may be related. Hence most of the sequences in the databases are related to one (and suitably only one) of the daughter sequences. This is one of the main foundations of the present invention. By comparing a very large number of mtDNA sequences it has been found that almost all the humans in Europe whose mtDNA sequences have been analysed were directly descended, through the maternal line, to just a few (probably 7) women all of whom lived thousand of years ago. This inescapable yet surprising fact that most people are related to each other through our mothers, and are ultimately only related to a few females, allows potential ancestry to be analysed by comparing mtDNA sequences.

Each daughter sequence may have one or more defining or different mutations.

Each daughter sequence (or defined mutation) may thus determine a different clade (or clan or group). For example, mutations at the following positions may define a daughter sequence or clade (a letter denoting the daughter sequence or clade is placed first, followed by positions of the defining mutations, and these are reflected in Figures 1 and 2): V is 298; J is 126 and 069; T is 126 and 294; K is 224 and 311; U is 270; and I, X and W all have 223. It will thus be seen that most clades or daughter sequences are defined by only one or two mutations, in other words mutations at either one or two positions. Thus, the clade to which the sample sequence belongs can be determined by detecting whether the sample sequence has one or more of the mutations present in the daughter sequence.

The reference and/or daughter sequence(s) are preferably from dead humans, for example they may be taken from fossil samples. Many (e.g at least 80%, 90% or 95%) of the comparison sequences, however, may be from humans that are still alive. The total number of sequences may be at least 1,000, for example 5,000 and preferably at least 10,000. Preferably there is only one reference sequence (for example the sequence in Figure 1).

The length of the sequences compared is preferably at least 50, at least 100, at least 200 or 350, and preferably at least 400 bases (or nucleotides). However, the length is preferably less than 1,000, such as less than 800. Preferably the sequences are compared over a length of 200 to 600, such as 300 to 500, bases.

Preferably the sequences that are used in the comparison are from the same region of mitochondrial DNA. There are about 16,600 bases in mtDNA (the exact number is thought to be 16,589). It is preferred that sequences are compared over a length which is the same as the length of the sequences. Comparisons are preferably made between hypervariable sequences, such as HVS I and/or HVS II, although the former is preferred.

A seventh aspect of the invention relates to a computer-based or aided method for conducting a method of the first aspect. The comparison in (b) can be performed using a database. The method may additionally involve the (e.g. computer) generating and/or displaying a gene tree. This aspect includes a method of conducting any of the methods of the first aspect where one or more steps are computer implemented, for example the comparison in step (b).

A eighth aspect of the invention relates to a computer program, for example present on a computer readable storage medium, which is capable of conducting the

comparison of the first, second, third or fourth aspect. The computer program may comprise program code means for conducting the method according to the first aspect. This aspect includes a computer program storage medium readable by a computing system and encoding a computer program of instructions for executing a computer process, the computer process being a method of the first aspect. The computer program may be present on a data carrier, and therefore this aspect includes a data carrier possessing the computer program.

Thus, this aspect of the invention includes a computer program comprising program code means for determining probable ancestry by:

(a) comparing an input mitochondrial DNA sequence from a human (the "sample sequence") with a multiplicity of stored mtDNA sequences ("comparative sequences") each of which is from a human different from the human having the sample sequence; and

(b) providing, on the basis of the comparison, an indication that the human having the sample sequence is a probable ancestor of a human female having a related comparison sequence.

A ninth aspect of the invention relates to a computer system (or an apparatus for use in a computing system) comprising means for:

(a) receiving or accepting an mtDNA sequence (the "sample sequence");

(b) means for comparing the sample sequence with a multiplicity of mtDNA sequences ("comparison sequences"), for example as specified in the first aspect; and

(c) a means for generating from the comparison the desired indication, correlation or determination as specified in any of any of the first aspects.

This aspect therefore includes a computer system programme to perform steps (b) and (c) of the first and second aspects, steps (b), (c) and (d) of the third aspect, or steps (a) and (b) of the fourth aspect.

The system may also comprise means for outputting or displaying a gene tree (for example showing the position in the tree of the sample sequence) or providing the clade of the individual.

The above computer implemented steps in another implementation of the invention are provided as an article of manufacture, i.e. a computer program being storage

medium readable and containing a computer process for performing the above described steps.

The computer system may have one or more modules for implementing each, or more than one, of the steps as described above.

The embodiment of the invention described herein can be implemented as logical operations in a computer system. The logical operations of the present invention can be implemented either as a sequence of computer implemented steps running on a computer system and/or as interconnected machine modules within a computer system. The method implementation on machine implementation is a matter of choice dependant on the performance requirements of the computer system implementing the invention. Accordingly, the logical operation making up the embodiment of the invention described herein are referred to variously as operations, steps or modules.

Preferred features and characteristics of one aspect of the invention are applicable for any other aspect *mutatis mutandis*.

The present invention will now be described by way of illustration with reference to the accompanying drawings and Examples, which are provided merely for illustration and are not to be construed as limiting:

Figure 1 is a gene tree showing a reference mtDNA sequence and the relationship of comparison mtDNA sequences by virtue of mutations with respect to the reference sequence; and

Figure 2 is a gene tree similar to Figure 1, showing a sample DNA sequence, from a human individual and the location of that sequence within the gene tree, thereby indicating probable ancestry.

EXAMPLES

Example 1

The research conducted has resulted in the identification of just seven women, who lived thousands of years ago, from whom almost everyone in Europe is directly descended through the maternal line. It thus seems that an unescapable fact is that the current human European population are all related to each other through our mothers. By using DNA comparisons and identifying mutations it has been possible to calculate how

long ago these seven women lived, and even estimate their geographical origin. It has been established that they survived and each had daughters, starting off lines of descent that reach in a number of maternal line right down to each individual alive today.

A tissue or fluid sample is taken from a human whose ancestry is to be determined. This is from a cheek sample, obtained by lightly scraping the inside of a cheek with a nylon brush, to obtain cheek cells. These cells have their mitochondrial DNA removed by using known chemical and mechanical techniques. Using DNA sequences that are commercially available, the entire mtDNA sequence was determined, approximately 16,600 bases. A particular section, HVS I, was identified, and this sequence (400 bases) chosen for comparison with sequences from the same region (HVS I) of mtDNA from other sources whose sequences are held on a database. This database was produced by analysing approximately 12,000 sequences from volunteers living in various parts of Europe and employs VARIANTS software. The same technique was applied to generate all of these sequences, namely taking a sample from the human, extracting the mtDNA and sequencing the DNA. The database also contains mtDNA sequences from dead humans, in particular from fossil samples.

Mitochondrial DNA is chosen because it is passed on only through the egg. This is unlike most DNA which is passed on to the next generation in sperm and eggs. This means that an individual has inherited their mtDNA only from their mother. In turn, their mother has inherited the mtDNA from their mother, who inherited it from their mother, and so on. Hence mtDNA can provide a direct and undiluted link to maternal ancestors.

DNA is copied when cells divide. This is usually an extremely accurate system and mistakes occur very rarely. These mistakes are known as mutations. They may have adverse effects if the mutation occurs in a gene responsible for an important cell function (for example giving rise to severe inherited diseases like cystic fibrosis or muscular dystrophy) although many mutations can occur without having any effect at all. Mutations in mtDNA that have been analysed here have been found to have no consequences (adverse or otherwise). In the past it has been calculated that, on average, one mutation occurs in mtDNA average every 20,000 years. This may sound infrequent, but in fact mtDNA mutates at a greater rate than DNA from other parts of the cell. It is therefore an ideal choice for studying mutations and using these to determine lineage.

The sequences in the database (comparison sequences) are correlated in order to

determine their differences, or the mutations that lead from one sequence to another. In this fashion a gene tree can be built up, as shown in Figure 1. The reference sequence is shown as the most central node, with four lines emanating therefrom, two of which are labelled 12308 and 00073. The mitochondrial DNA reference sequence shown at the bottom of Figure 1 corresponds to this node. The node directly above it and to the left, labelled 223 in Figure 1 and possessing a dotted line and embolded in Figure 2, indicates lineage to mtDNA sequences outside Europe, i.e. to the rest of the world. Figure 2 is in effect the same gene tree as Figure 1 except some details are omitted and a sample sequence (from an alive human) is shown together with its position in the gene tree. Figures 1 and 2 only concern European mtDNA sequences. The nodes emanating from the central node, depicted by the numbers 12308, 00073 and 07028 all have the same sequence as the reference sequence (within the region being compared); these numbers merely depict mutations outside the portion of the reference sequence under study. These unshaded spheres (nodes) all represent sequences from dead humans, i.e. ancestors.

Figure 1 thus shows the lineage by mutation. Starting from the reference sequence, and moving in a south-westerly direction, the first line is labelled 126, which means that the node at the end of this line depicts a sequence that is different from the reference sequence by a mutation at position 126. Moving further downwards to the node marked J, there is a line marked 069 which indicates an additional mutation at position 69. This means to proceed from the reference sequence to the sequence represented by J, there are two mutations, namely at positions 126 and 69. In this way all the comparison sequences in the database can be related to each other by means of this gene tree. These numbers refer to the (control or study region of) reference sequence (400 bases) and correspond to the positions in the HVS I region minus 16,000. Thus, the first A of the reference sequence shown is for convenience position 1 (1601 using known mtDNA nomenclature).

The numbers indicate the position of the mutation: almost all mutations are transitions (that is to say exchange of one purine base for another, and so are C→T, T→C, A→G or G→A). There are only a few transversion mutations (A→C or T etc) and these are shown separately with the nature of the mutation (e.g. see top left corner of Figure 2, mutation 114 C→A).

In Figures 1 and 2 each of the nodes or spheres represents a different DNA

sequence. The size of the sphere is proportional to the number of people who have that particular sequence. For example, sphere H, which is the largest sphere, represents the most frequent mtDNA sequence found in Europe. It is also the reference sequence that has been used to compare other sequences in this Example. The lines linking the spheres thus trace the maternal evolutionary pathways, and their length depends on the number of changes between the sequences. Each line usually indicates a single change of mutation in the 400 nucleotide sequence. This chart, or "gene tree", was prepared using over 12,000 sequences from people living in Europe and other continents.

The five white spheres (or nodes) in the centre of the Figures represent sequences that have never been found in any living person, but in reconstructions it has been found that humans having these sequences must have existed.

The sample sequence shown at the bottom of Figure 2 is a sequence from an individual whose ancestry is to be determined. This sequence is then compared with the reference sequence and the differences (mutations) determined. In this case, when comparing with the reference sequence (Figure 1) three mutations were identified, at positions 126, 292 and 294. The bases at these three positions are shown in bold in the sample DNA sequence in Figure 2. Therefore, by comparing the sample mtDNA sequence in Figure 2 with the reference mtDNA sequence in Figure 1, notices three mutations (C for T at position 126, T for C at position 292 and T for C at position 294). With the knowledge of these mutations, one can follow the ancestral line from the central node in Figure 1 via lines labelled 126, 294 and then on to 292, and the position in the gene tree is thereby determined (shown by the star in Figure 2). This shows that the individual belongs to the group or clade designated T, and is an ancestor of the female having the sequence labelled T.

The seven women from whom 99% of the sequences in the database are related are shown in Figure 1 and are labelled H, V, J, T, K, U and X (X stands for the three groups I, W and X). These constitute the daughter sequences. Thus all the sequences on the database can be split into seven different groups, clades or clans, according to the mutations between sequences. The members of each group, clade or clan are each descended from a single woman through the maternal line.

The invention can then be used to obtain information from the sample mtDNA sequence once information is known about the seven women. From the fossils and other

information it can be determined how long ago that women lived, her group or clan or race, her geographical origin, and the geographical region in which she lived.

Example 2: Calculation of the age of an ancestral mtDNA sequence in Europe

In order to study the timing and extent of demographic expansions using mitochondrial DNA variation one needs to identify founder lineages in potential source populations. The present work concerns a method for determining the age of a mitochondrial sequence in Europe.

European mtDNA can be broken into about a dozen major clusters, referred to as H, V, U3, U4, U5, K, J, T, I, W and X. The RFLP haplogroup U subsumes both haplogroup K and a number of other clusters now referred to as U3, U4 and U5, and includes several clusters found only rarely in Europe, including U1, U2, U6 and U7, which occur more commonly in North Africa and the Near East.

The Near East was taken to include the whole of Turkey, the Fertile Crescent from Israel to western Iran, and the whole of the Arabian peninsula. The lower Nile (Egypt and northern Sudan) was also included, since this region is often treated historically with the Near East. HVS I sequence data show a large proportion of Near Eastern mtDNAs have penetrated the Nile Valley, where they co-exist with sub-Saharan African mtDNAs. The Near Eastern populations analysed for sequence variation in the first hypervariable segment (HVS I) of the mitochondrial control region were therefore as follows: 80 Nubians, 68 Egyptians, 29 Bedouin, 43 Yemenites, 112 Iraquis from four regions of Iraq, 11 Iranians from both Iran and Germany, 67 Syrians, 45 east Jordanians (from the Dead Sea region), 100 west Jordanians (from the Amman region), 117 Israeli Palestinians, 45 Israeli Druze, 221 Turks from Turkey, 53 Kurds from eastern Turkey, 67 Syrians from Damascus, and 190 Armenians from Armenia.

The European populations were analysed on the basis of the palaeoclimatological model (Gamble C (1986) The Palaeolithic settlement of Europe, Cambridge University Press, Cambridge, pp471). South-east: 142 Bulgarians. Mediterranean-east: 69 Greeks from Thessaloniki; 62 Sarakatsani from Epirus. Mediterranean-central: 49 Italians from Tuscany and 47 from Rome; 89 Sicilians;115 Sardinians. Mediterranean-west: 54 Portuguese; 72 Spaniards; 94 Galicians; 156 Basques from northern Spain. Alpine: 70 Swiss; 49 South Germans from Bavaria; 100 Austrians. North-central: 34 Poles ; 84

Czechs; 109 Germans; 39 Danes. Scandinavia: 32 Swedes; 232 Norwegians; 56 Icelanders. North-west: 100 British; 94 Cornwall; 92 from Wales; 101 from western Ireland. North-east: 195 Finns; 48 Estonians; 34 Volga-Finns; 26 Russians. Additional data from mtDNAs of ambiguous haplogroup classification were obtained by screening for diagnostic markers. This screening mainly involved haplogroups H (7025 *Alu*I), HV (14766 or 00073) and U (12308 *Hin*fI).

Several rounds of founder analysis and dating were performed. Each identified candidate European found haplotypes on the basis of matching with Near Eastern haplotypes, and estimated the age of the candidate founder in Europe by measuring the diversity accumulated along the lineage since the (founder event) in Europe.

We first used the data raw, assigning founder status to every matching sequence type. Founder candidates were listed, reduced median networks (Bandelt *et al*, (1995) Mitochondrial portraits of human populations using median networks. Genetics 141:743-753) constructed for the cluster radiating from each founder sequence, and their diversity estimated using the statistic ρ. This value was converted to a point estimate of the age using a mutation rate of 1 transition per 20,180 years (Forster P, Harding R, Torroni A, Bandelt H-J (1996) Origin and evolution of Native American mtDNA variation: a reappraisal. Am J. Hum Genet 59:935-945), which approximates to the standard rate used for HVS I (Macaulay VA, Richards MB, Forster P, Bendall KA, Watson E, Sykes BC, Bandelt H-J (1997) mtDNA mutation rates, Am J. Hum Genet 61:983-986). Central 95% credible regions were calculated using the method (Berger JO (1985) Statistical Decision Theory and Bayesian Analysis: Springer-Verlag, New York) to allow for sampling variation. In some cases, additional (unsampled) founder types were identified as nodes in the phylogenetic networks. For example, a case may occur in which a Near Eastern type leads to an unsampled branching node, leading to both Near Eastern and European derivatives. This unsampled node was necessarily considered as a founder candidate in both crude and refined analyses.

We then attempted to correct for the high levels of recurrent gene flow suggested by the data (see below) by employing additional criteria for a candidate found lineage beyond it merely matching a Near Eastern sequence. The refined analysis began with matching sequences, as did the crude analysis. However, to shield against sporadic examples of back-migration into the Near East from Europe (and to some extent also

back-mutation) the three criteria were developed for the identification of founders amongst the shared lineages:

(1) the matching sequence must occur in, or have legitimate derivatives in, at least two Near Eastern populations;

(2) the matching sequence must have legitimate derivatives in at least one Near Eastern population;

(3) the matching sequence must have legitimate derivatives in at least two Near Eastern populations.

These criteria are designed to identify lineages that most likely evolved in the Near East region, rather than having entered by migration in the more recent past. Legitimate derivatives must connect to the founder candidate via Near Eastern sequence types, and not by sequence types found only in Europe. The details of these criteria were the result of trial and error attempts to eliminate U5 and V entirely as Near Eastern founders. At the same time the criteria have to be sufficiently slack to allow such a sequence as the root type of cluster J1a to be identified as a founder. It should clearly be classified as such, because it has diversified considerably in the Near East and very little in Europe, but the root type itself occurs only once in the Near Eastern sample.

In fact, only criterion (3) achieved the elimination of the root types of both and U5 and V as possible founders. To guard against the possibility that such a criterion would be too severe, however, criteria (1) and (2) were also employed, and the results calculated using a set of consensus founders. Criterion (1) identified 70 founders, (2) identified 73, and (3) identified only 52. 57 were identified by consensus.

European population divergences from the Near East were estimated using the ρ statistic. Divergences for each geographic region were estimated against the list of matching sequence types. The average divergence of the region in question against the divergence of the Near East from the founder list was used as an estimate of population separation times (see Barbujani G, Bertorelle G, Chikhi L (1998) Evidence for Paleolithic and Neolithic gene flow in Europe, AM J. Hum Genet 62:488-4912) for mtDNA and (Chikhi L, Destro-Bisol, Bertorelle G, Pascali, Barbujani G (1999) Clines of nuclear DNA markers suggest a largely Neolithic ancestry of the European gene pool, Proceedings of the National Academy of Sciences USA 95:9053-9058) for Y-chromosome microsatellites. These ages are referred to as "effective divergence times" since on a model of founder

effects and substantial subsequent gene flow (rather than population splits and subsequent isolation) they are meaningless as dates. The analysis was then repeated using the refined set of founders and the effective divergence times estimates recalculated.

Table 1 shows frequencies and age estimates of the main mitochondrial clusters in the Near East and Europe. Central 95% credible regions on sampling error are also given for both frequencies and age estimates. These ranges warn against taking the point estimates too literally.

The age estimates date approximately the time of origin of each cluster. Since these clusters tend to be restricted to Europe and the Near East, they are likely to have originated either in one or the other region, and expanded into the other (at least, on a simple "island" model which regards Europe and the Near East as distinct populations). Therefore, the older of the two age estimates should indicate the age of the cluster, and its region of origin. A founder analysis is then necessary to date the expansion into the population where it is younger and the crude estimate for the younger population becomes, in itself, rather meaningless.

TABLE 1

Ages of major clusters in Europe

cluster	*n* Europe	% Europe	95% range	*m* Europe	age Europe	95% range Europe
U1	14	0.6	0.3-1.0	37	54800	38800-73500
U2	13	0.6	0.3-0.9	20	32600	20200-47900
U3	26	1.1	0.7-1.6	19	14800	8900-22100
U4	70	2.9	2.3-3.6	71	20800	16300-25800
U5	225	8.1	9.2-10.4	533	47900	43900-52000
K	148	6.1	5.2-7.1	116	16000	13200-19000
H	1132	46.4	44.4-48.3	1090	19500	18300-20600
V	76	3.2	2.5-3.9	116	13400	10600-16500
J	233	9.6	8.4-10.8	276	9400	7600-11200
T	206	8.5	7.4-9.6	468	17500	14500-20500
I	48	2.0	1.5-2.6	77	32800	25900-40500
W	49	2.0	1.5-2.6	44	18500	13500-24300
X	39	1.6	1.2-2.2	41	21700	15700-28800

CLAIMS

1. A method of determining probable ancestry, the method comprising:

(a) providing a mitochondrial DNA sequence from a human (the "sample sequence");

(b) comparing the sample sequence with a multiplicity of mtDNA sequences ("comparison sequences") each of which is from a human different from the human having the sample sequence; and

(c) providing, on the basis of the comparison, an indication that the human having the sample sequence is a probable ancestor of a human female having a related comparison sequence.

2. A method according to claim 1 wherein the human having the sample sequence is alive and/or one or more of the comparison sequence(s) is from a dead human.

3. A method according to claim 1 or 2 which additionally comprises taking a tissue or fluid sample from a human, extracting mtDNA from that sample, and determining the sequence of at least part of that mtDNA.

4. A method according to any preceding claim wherein the comparison in (b) comprises determining the closest (comparison) sequence to the sample sequence.

5. A method according to claim 4 wherein the closest sequence belongs to a dead human and/or represents the closest (female) ancestor.

6. A method according to any preceding claim which additionally comprises correlating a comparison sequence to the sample sequence or another comparison sequence by comparing the sequences to a reference sequence, determining the differences between the sequences (mutations) and then correlating the sequences that possess the same mutations.

7. A method according to any preceding claim which comprises generating correlated comparison sequences on the basis of their mutations and/or placing the sample sequence on a gene tree or diagram.

8. A method according to any preceding claim wherein some of the comparison sequences are daughter sequences, which have from 1 to 4 mutations from the reference sequence, and wherein at least 90% of the comparison sequences possess one or more of the same mutations as the daughter sequences.

9. A method according to any preceding claim wherein at least 95% of the

comparison sequences are all related to one of the daughter sequences.

10. A method according to claim 9 wherein there is one reference sequence and/or the number of daughter sequences is from 3 to 10.

11. A method according to any preceding claim which further comprises determining one single human female only as an ancestor.

12. A method according to claim 11 wherein the ancestor died at least 5,000 years ago.

13. A method according to claim 11 or 12 wherein the human female ancestor is determined from 3 to 10 human female ancestors, all of which died at least 10,000 years ago.

14. A method of determining probable ancestry or obtaining ancestral information, or determining the clade of an individual, the method comprising:

(a) providing an mtDNA sequence from an individual (the "sample sequence");

(b) comparing the sample sequence with a reference mtDNA sequence from a human and determining the differences (mutations) between the sample and reference sequences;

(c) using the nature of the mutation(s) to correlate the sample sequence with an mtDNA sequence from a human other than the individual (the "comparison sequence") where the sample and comparison sequences both have at least one mutation in common; and

(d) providing, on the basis of the correlation between the sample sequence and the comparison sequence, a determination of the clade of the individual, a probable ancestor or information about the human female ancestor.

15. A database comprising at least 10 mtDNA sequences, the database being structured so that each sequence is correlated with one or more pieces of information concerning a human female ancestor having that DNA sequence.

16. A database according to claim 15 which has at least 1,000 sequences.

17. A database according to claim 15 or 16 which has one reference sequence and from 3 to 10 daughter sequences, each daughter sequence having from 1 to 4 mutations from the reference sequence.

18. A database according to claim 17 wherein at least 90% of all the sequences

are related to a daughter sequence.

19. The use of a database according to any of claims 15 to 18 in the determination of a probable human female ancestor from a mitochondrial DNA sequence.

20. A computer program comprising program code means for determining probable ancestry by:

(a) comparing an input mitochondrial DNA sequence from a human (the "sample sequence") with a multiplicity of stored mtDNA sequences ("comparison sequences") each of which is from a human different from the human having the sample sequence; and

(b) providing, on the basis of the comparison, an indication that the human having the sample sequence is a probable ancestor of a human female having a related comparison sequence.

21. A computer program according to claim 20 wherein the human having the sample sequence is alive and/or one or more of the comparison sequence(s) is from a dead human.

22. A computer program according to claim 20 or 21 wherein the comparison comprises determining the closest (comparison) sequence to the sample sequence.

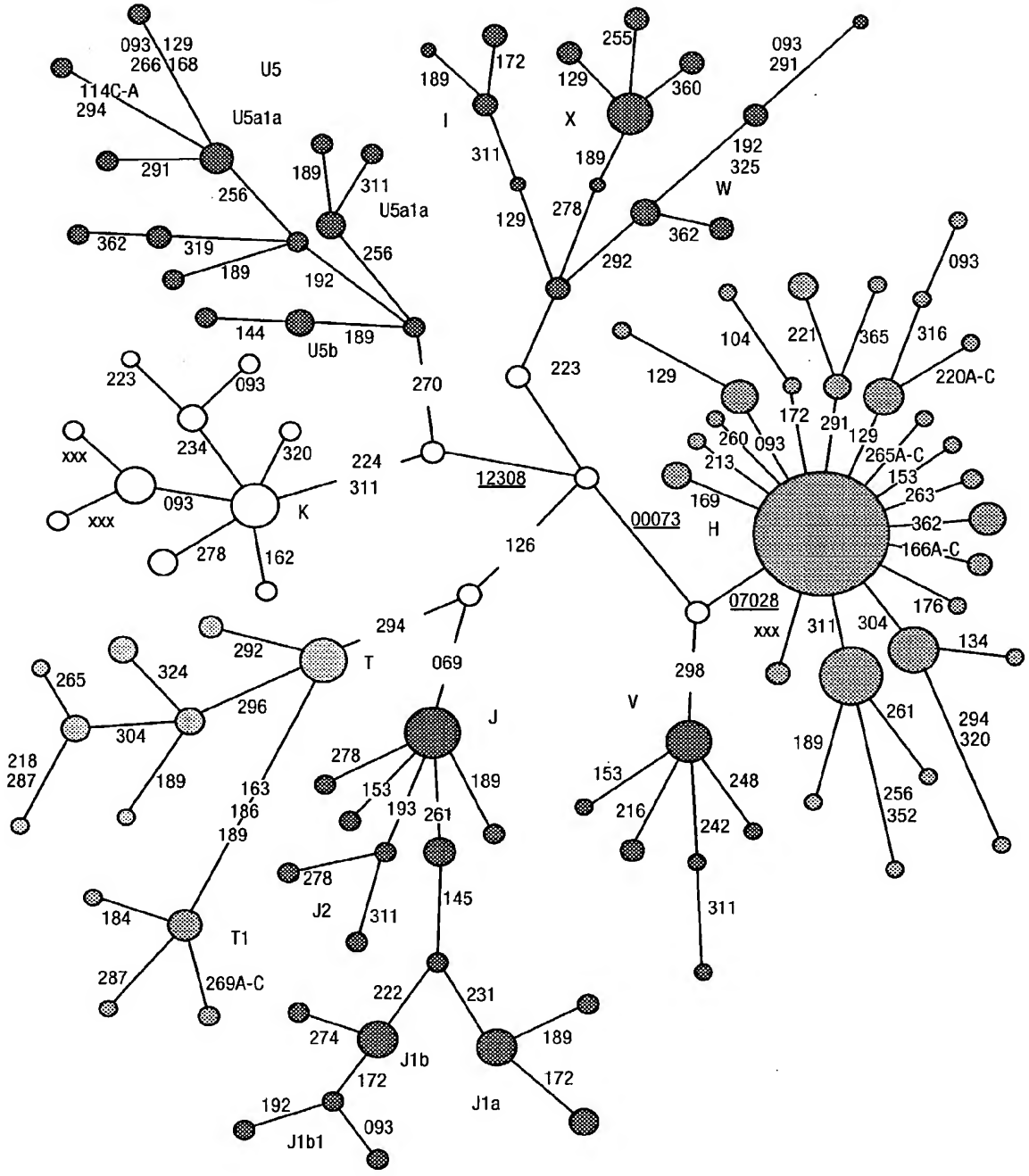
23. A computer program according to claim 22 wherein the closest sequence belongs to a dead human and/or represents the closest (female) ancestor.

24. A computer system programmed to perform steps (b) and (c) of claim 1 and/or steps (b), (c) and (d) of claim 14.

25. A method of determining probable ancestry substantially as herein described with reference to the accompanying Examples.

26. A computer program capable of performing steps (b) and (c) of claim 1 or (a) and (b) of claim 20 substantially as herein described with reference to the Examples.

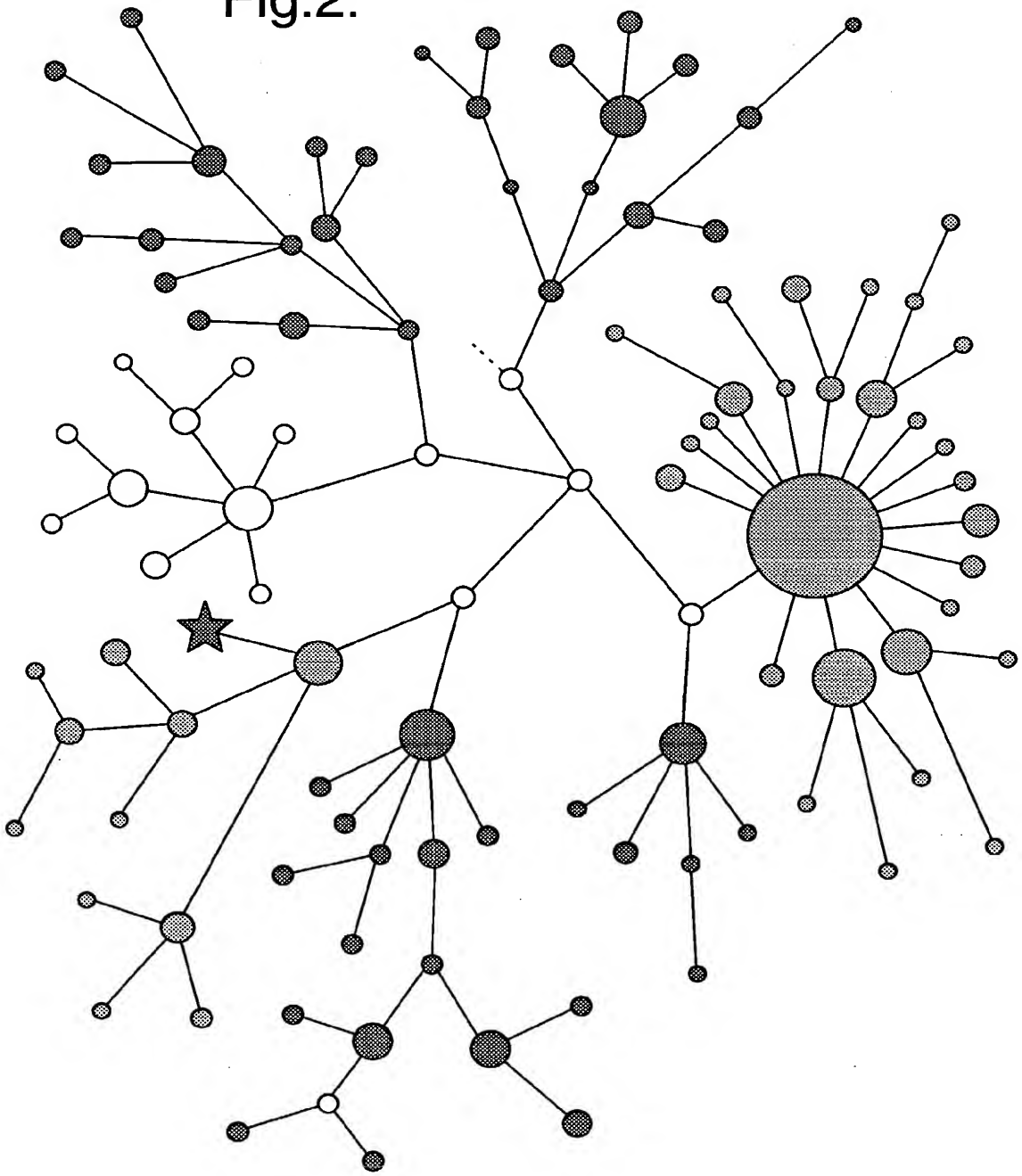
Fig.1.



mitochondrial DNA sequence

ATTCTAATTT AAACTATTCT CTGTTCTTTC ATGGGGAAGC AGATTTGGGT ACCACCCAAG TATTGACTCA CCCATCAACA ACCGCTATGT ATTTCGTACA
TTACTGCCAG CCACCATGAA TATTGTACGG TACCATAAAT ACTTGACCAC CTGTAGTACA TAAAAACCCA ATCCACATCA AAACCCCCTC CCCATGCTTA
CAAGCAAGTA CAGCAATCAA CCCCCAACTA TCACACATCA ACTGCAACTC CAAAGCCACC CCTCACCCAC TAGGATACCA ACAAACCTAC CCACCCTTAA
CAGTACATAG TACATAAAGC CATTTACCGT ACATAGCACA TTACAGTCAA ATCCCTTCTC GTCCCCATGG ATGACCCCCC TCAGATAGGG GTCCCTTGAC

Fig.2.



DNA sequence

ATTCTAATTT AAACTATTCT CTGTTCTTTC ATGGGGAAGC AGATTTGGGT ACCACCCAAG TATTGACTCA CCCATCAACA ACCGCTATGT ATTTCGTACA
TTACTGCCAG CCACCATGAA TATTGCACGG TACCATAAAT ACTTGACCAC CTGTAGTACA TAAAAACCCA ATCCACATCA AAACCCCCTC CCCATGCTTA
CAAGCAAGTA CAGCAATCAA CCCCCAACTA TCACACATCA ACTGCAACTC CAAAGCCACC CCTCACCCAC TAGGATACCA ACAAACCTAC CCACCCTTAA
CAGTACATAG TACATAAAGC CATTTACCGT ACATAGCACA TTACAGTCAA ATCCCTTCTC GTCCCCATGG ATGACCCCCC TCAGATAGGG GTCCCTTGAC